



IA et bioinformatique

Jacques Nicolas

► To cite this version:

Jacques Nicolas. IA et bioinformatique. Sébastien Konieczny; Henri Prade. L'intelligence artificielle: De quoi s'agit-il vraiment ?, Cépaduès, pp.101, 2020, 9782364938502. hal-03127545

HAL Id: hal-03127545

<https://inria.hal.science/hal-03127545>

Submitted on 1 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

IA et bioinformatique

Jacques Nicolas
Inria Rennes
jacques.nicolas@inria.fr

Les liens entre intelligence artificielle et bioinformatique remontent à l'origine des disciplines : Joshua Lederberg, professeur de génétique à Stanford, a établi des échanges dès la fin des années 1960 avec Edward Feigenbaum et Bruce Buchanan, deux chercheurs en intelligence artificielle. Un des premiers résultats de ces efforts, le système expert *DENDRAL*, identifiait les composés organiques à partir de données de spectrométrie de masse et d'une base de règles exploitant des connaissances spécifiques au domaine. Puis, parmi les premières applications d'apprentissage automatique, Meta-DENDRAL a appris les règles nécessaires à DENDRAL à partir de paires (spectres, structures).

Les raisons de cette convergence d'intérêt tiennent probablement à deux facteurs :

- D'abord, les organismes vivants sont une source d'inspiration pour l'intelligence artificielle pour mettre au point des méthodes robustes par rapport au traitement de données du monde réel : la tolérance aux données imprécises et incertaines et la capacité d'auto-adaptation et d'apprentissage se rencontrent dans de nombreux exemples en biologie. Les méthodes correspondantes se regroupent parfois sous le terme de "*soft computing*" (réseaux neuronaux, algorithmes évolutifs, logiques de l'incertain, techniques d'optimisation...) ;
- Ensuite, la biologie est une science de la connaissance par excellence. On y observe un jeu complexe d'influences causales contextuelles et hiérarchiquement organisées, en lien avec des fonctions biologiques ciblées. Qu'il s'agisse de santé, d'agronomie ou d'étude de la biodiversité, la biologie cherche à donner un sens à cet énorme ensemble de relations. Par ses questionnements sur la représentation des connaissances et l'automatisation du raisonnement, l'intelligence artificielle permet d'affronter le double challenge de l'intégration des observations et des connaissances et de l'exploitation intensive de ces informations.

L'un des premiers enjeux de l'intelligence artificielle en bioinformatique est le traitement des connaissances. En biologie moléculaire, le flux à haut débit de données de multiples dispositifs d'observation combiné aux résultats d'analyses est actuellement recueilli dans des centaines de bases de données. La maintenance puis l'intégration de ces bases de différents niveaux de qualité est extrêmement complexe. L'enjeu est la transition d'un ensemble d'îlots d'expertise mal formalisés et partiellement erronés, vers une vision cohérente et unifiée de connaissances interdépendantes. L'intelligence artificielle apporte des outils essentiels pour accompagner cette transition et conférer du sens aux données. Ceci comprend la création d'ontologies, qui structurent un domaine de façon contrôlée, en identifiant de façon non ambiguë les entités, leurs propriétés et relations avec les autres entités, ainsi que la représentation de graphes de données hétérogènes, en particulier via des formalismes logiques. Différentes formes de raisonnement automatisé sur les

connaissances sont étudiées, qu'il s'agisse de traquer les incohérences, d'extraire des causalités cachées ou d'interrogation intelligente du contenu.

Un vaste ensemble d'informations en biologie et santé résiste aux efforts pour normaliser et exploiter les données : les ressources textuelles (comptes-rendus hospitaliers, articles de revues). Le savoir-faire de l'intelligence artificielle en analyse de textes en langage naturel et en fouille de textes permet, non pas de tout automatiser, mais d'assister les curateurs, personnages clés de la formalisation des connaissances, en filtrant, en mettant en forme l'information et en proposant et vérifiant des associations d'entités. On peut ainsi chercher à distinguer dans un article les parties traitant du dispositif ou des résultats, les faits accrédités ou rejetés par une expérience, les hypothèses... Signalons pour terminer le passionnant nouveau domaine de la compréhension automatique de figures, source très riche d'informations.

Un autre axe important est celui de la prédiction, à des fins de diagnostic, de pronostic ou de sélection d'hypothèses les plus intéressantes à tester. L'apprentissage automatique est au cœur de ces travaux en bioinformatique. On peut vouloir prédire les sites actifs ou la conformation tridimensionnelle d'une protéine, les interactions entre différentes molécules ou encore les facteurs génétiques déterminant un développement, une maladie, ou un fonctionnement particulier. L'application à cette fin de méthodes connexionnistes s'est récemment énormément développée mais on a également recours à des méthodes plus explicites comme le raisonnement à partir de cas, l'apprentissage d'arbres de décision, de grammaires formelles ou de programmes logiques. En effet, il ne s'agit pas seulement d'obtenir un prédicteur efficace, mais aussi de comprendre et d'éclairer un processus de décision qui reste manuel car impliquant des coûts importants, en termes de santé ou d'expérimentation. L'aide à l'expérimentation et à la découverte reste un thème très important de l'intelligence artificielle en bioinformatique.

Enfin il nous faut citer un dernier axe important, la résolution de problèmes combinatoires. En effet, l'approche mathématique analytique n'est pas toujours applicable en biologie où les équations régissant un système peuvent être extrêmement complexes. Comprendre le fonctionnement ou la structure d'un système qui mélange une hiérarchie d'aspects discrets et continus suppose l'emploi de méthodes symboliques et hybrides, en particulier les systèmes de contraintes et les solveurs logiques. Une caractéristique fondamentale des problèmes en bioinformatique est l'existence de solutions alternatives multiples aux problèmes d'optimisation, un domaine que l'intelligence artificielle excelle à traiter notamment par l'exploration intelligente de grands espaces d'hypothèses et la gestion d'ensembles de préférences. Un autre aspect est celui de la gestion de faits potentiellement contradictoires que seule une approche basée sur la connaissance permet de gérer proprement. Par exemple, l'étude de l'évolution peut être abordée par un va-et-vient constructif homme-machine pour la construction d'arbres d'espèces reflétant l'évolution observée des caractères mais aussi les hypothèses effectuées sur le processus d'évolution lui-même. La vérification de modèles, l'usage des logiques temporelles pour vérifier certaines propriétés, ou la production automatique de contre-exemples, sont autant d'aides précieuses pour amplifier les possibilités de raisonnement du scientifique ou du praticien.